DOCUMENT RESUME

ED 244 999 TM 840 323

AUTHOR Harvill, Leo M.
TITLE Measuring the Test-Wiseness of Medical Students.
PUB DATE 24 Apr 84
NOTE 9p.; Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Higher Education; Measurement Techniques; *Medical Students; *Multiple Choice Tests; Test Coaching; Test Construction; Test Items; *Test Reliability; *Test Validity; *Test Wiseness
IDENTIFIERS *Measure of Obscure Medical Information

ABSTRACT

The objectives for this study were to: (1) develop a valid, reliable measure of test-wiseness with equivalent forms for use with students in the health sciences; and (2) determine the level of test-wiseness of entering medical students. The test-wiseness areas included in this study were: similar options, umbrella term, item give-away, convergence principle, length of correct option, and stem-option cue. The content for the test items was based on obscure, factual medical information taken from "Dorland's Illustrated Medical Dictionary." Each of the two test forms contained 24 four-choice multiple-choice items. The test, entitled Measure of Obscure Medical Information, was administered to 15 students enrolled in a workshop on test-taking skills and 39 other entering medical students. Although the test is still in the early stages of development, there is some evidence of both internal consistency and alternate form reliability. Some degree of validity for measuring test-wiseness is indicated by the significant gain made on the test by students instructed in test-taking skills. The test seems appropriate for use with health science students because of the nature of the content and the level of performance on the test by medical students. (BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

L. M. Harvill

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

NCME/AERA Annual Meeting
New Orleans, Louisiana
April 24, 1984

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED244999

Measuring the Test-Wiseness of Medical Students

Leo M. Harvill, Ph.D.
Quillen-Dishner College of Medicine
East Tennessee State University

Test-wiseness (TW) has been defined by Millman, Bishop and Ebel (1965) as "a subject's capacity to utilize the characteristics and formats of the test and/or the test-taking situation to receive a high score." A number of measures of TW have been constructed to assess several aspects of TW; these are discussed in Sarnacki's (1979) excellent review article. These various instruments have measured the following TW principles: 1) stem-option, 2) absurd options, 3) specific determiners, 4) precise options, 5) longer length options, 6) grammatical, 7) item give-away, 8) placement of correct option, 9) similar options and 10) opposites. In addition, Smith (1982) developed an instrument to measure what he called the convergence principle. Obviously, there are many aspects of TW which are not included above, such as appropriate use of testing time and appropriate guessing strategies.

There appears to be a need for a TW instrument with equivalent forms for use with health science students. A good measure of TW would be of value in determining what impact this set of skills has on test performance and grades in health science schools. Such an instrument would also be helpful for measuring the effectiveness of instructional programs designed to improve TW. Sarnacki (1981) suggested "the necessity of a better, more discriminating measure of TW for use with adults in general, and with medical undergraduates in particular."

The objectives for this study were to: 1) develop a valid, reliable measure of several aspects of TW with equivalent forms for use with students in the health sciences, 2) determine the level of several aspects of test-wiseness of entering medical students.

METHODS

The TW areas included in this study were: 1) similar options (all but one of the options are similar in meaning), 2) umbrella term (one of the options is an inclusive term which includes all other options), 3) item give-away (the correct answer to the item is revealed in another test item), 4) convergence principle (two or more options have one factor in common; two or more options have a second factor in common; select the option where the two factors converge), 5) length of correct option (the keyed option is significantly longer than the other options), 6) stem-option cue (there is an obvious resemblance between a word or phrase in the item stem and a word or phrase in an option).

The first four types of TW listed above would appear to fit under the broad heading of "elements independent of test constructor or test purpose" and the subheading of "deductive reasoning strategy" in the outline of test-wiseness principles provided by Millman, Bishop and Ebel (1965). However, the fourth type, convergence principle, clearly is related to the way individual writers construct items and could also fit under the broad heading of "elements dependent upon the

TM840 323

test constructor or purpose" in the outline mentioned above. The fifth and sixth areas included in this study fit clearly under that heading and the sub-heading, "cue-using strategy."

The six areas of TW were selected, in part, because each could lead the test-wise examinee to the correct answer and not just assist the examinee in eliminating one or more options as incorrect.

The content for the test items was based on obscure, factual medical information taken from Dorland's Illustrated Medical Dictionary (1974). The title for the two forms of the test is the Measure of Obscure Medical Information (MOOMI). Each of the two forms of the MOOMI contains 24 four-choice multiple-choice items in which the examinee is to select the one best answer. Each form has subtests of four items each for each of the six TW areas. An example of the test items measuring the convergence principle is:

A Scotch douche is

*a) alternating hot and cold water applied in a single jet stream.

b) cold water applied to the body in a single jet stream.

c) alternating hot and cold water applied in a fan-shaped spray.

d) warm water applied to the body in a single jet stream.

In an attempt to validate that the content was obscure or unknown to entering medical students, two faculty physicians (family physician and surgeon) and three medical students (sophomore, junior, senior) were asked to verify this on an item-by-item basis. The physicians agreed that the content of all 48 items would be unknown to entering students. The students found a total of three items which two of them believed entering students might know and two other items which only one student believed might be known to the entering student. On that basis, it was assumed that the test items did measure obscure information and that nearly all health science students could only arrive at the correct options by guessing or by the use of TW cues.

Eight judges (two faculty physicians, three medical educators, three medical students) were given the definitions of the six TW areas included in the test and the two forms of the test and were asked to categorize each item according to the TW principle it was measuring. They were not told how many of each item type to expect. This was a means of validiation of the six separate TW subtests within each test form. The only category which was unclear to some of the judges was the convergence principle; four judges misclassified or failed to classify more than half of these test items. In nearly all other cases, the judges categorized seven or eight of the eight items on the two forms correctly.

The test directions appear as follows:

Measure of Obscure Medical Information

DIRECTIONS: This test is a measure of your knowledge of actual, although little-known, facts in medicine. Answer every question; there is no penalty for guessing. There is no time limit. Mark your answers on the separate answer sheet; make no marks on the test copy.

Each time the test forms have been used, no time limit has been imposed. Nearly all entering medical students completed one test form (24 items) in 20 minutes or less.

The two test forms were first used in a 2-1/2 hour workshop covering effective test taking skills offered during a week-long orientation program the week prior to the opening of medical school. The elective workshop was taken by 15 of the 18 entering medical students involved in the pilot orientation program. One component of the workshop was concerned with the TW principles measured by MOOMI. Form A was used as a pretest and form B served as a posttest with no prior testing of the two test forms.

The two test forms were administered to 39 other entering medical students the following week during a three-hour testing session which also included the administration of several psychological tests. Twenty of the students completed form A first; nineteen started with form B. There was no time intervening between the two forms. The students in both testing situations received no other directions or explanations other than those written on the test form.

FINDINGS

Table 1 contains summary information for the two forms of the MOOMI. The mean values for the two forms are somewhat disparate (14.62 and 16.54). The standard errors of measurement (2.04 and 1.94) and K-R 20 reliability values (0.67 and 0.60) are similar for the two forms. The K-R 20 reliability value for the 48 item test (forms A and B combined) is .79.

Table 2 reports the alternate form reliability coefficients for the MOOMI. The subtest reliabilities range from 0.29 to 0.45 and are similar to the K-R 20 reliability values for the combined forms reported in Table 1. The alternate form reliability for the total test is 0.72. Using the Spearman-Brown formula, the reliability is predicted to be 0.84 for a test twice as long.

Intercorrelations among the subtest and total test scores for forms A and B combined are shown in Table 3. The intercorrelations among the eight item subtests range from -0.01 to 0.51 with a median value of 0.33. The items measuring the convergence principle have the lowest correlations with the other subtests.

The median biserial correlation between individual items and the four-item subtests for form A is 0.74 and 0.83 for form B. The lowest such biserial correlation is 0.43. The median biserial correlation between individual items and the twenty-four item tests is 0.51 for form A and 0.50 for form B. Seven of the 48 items had biserial correlations less than 0.20 with the total test score (form A or B).

Table 4 summarizes the data from the use of the MOOMI in the test taking skills workshop. Students made statistically significant gains from pretest to posttest on five of the six subtests and on the total test. The mean pretest score (form A) was 15.13 (63.0%) while the mean posttest score (form B) was 21.07 (87.8%). This finding must be tempered by the fact that form B is evidently not as difficult as form A but the total test gain would still be statistically significant after accounting for the difference in form difficulty.

At the conclusion of the workshop, students responded to form B and then categorized each item according to the area it was intended to measure. The correlation between the number of items correctly classified and the number of items correctly answered was 0.88. The mean number of items classified correctly was 20.13 (83.9%) while the mean number answered correctly was 21.07 (87.8%). The small difference can be attributed to guessing.

In an attempt to arrive at alternate forms with more equivalent mean scores, four items from form A were interchanged with four items from form B and the tests were rescored for the group of 39 medical students. Some of the analysis of the revised test forms is presented in Tables 5 and 6. Table 5 can be compared with Table 1 and Table 6 with Table 2. It can be seen that the means, standard deviations, standard errors of measurement, and Kuder-Richardson 20 reliability values are nearly identical for the two revised forms. The alternate form reliabilities for the two revised forms are similar to those for the two initial forms (see Table 2).

DISCUSSION

Although the test is still in the early stages of development, it appears that there is some evidence of both internal consistency and alternate form reliability. Some degree of validity for measuring TW is indicated by the significant gain made on the test by students specifically instructed in test taking skills. The test seems appropriate for use with health science students because of the nature of the content and the level of performance on the test by the medical students.

Entering medical students seem to have the ability to use TW principles. All of the mean scores for the four-item subtests reported in Table 1 are significantly above the chance or random guessing level except the convergence subtest on form A ($\bar{X} = 1.33$). They seem particularly adept at using the TW principles of item give-aways, stem-option cue, and umbrella term with mean scores for eight items all greater than 75 percent. They also did well on the test items dealing with the TW principle of similar options ($\bar{X} = 70\%$). These principles, except for the stem-option cue, deal with deductive reasoning strategies and are independent of the test constructor according to Millman, Bishop and Eble (1965). The performance of the medical students on these subtests indicates that they are able to use deductive reasoning in taking tests.

The performance of the entering medical students on the eight items measuring each of the TW principles of length of the keyed option and convergence was much poorer with mean scores of 44.5 percent and 37.1 percent, respectively. Over half (51.3%) of the students scored at or below the level expected by chance along on the eight convergence test items; only four students (10.3%) scored 75 percent or higher on those items. Appropriate use of both of these TW principles seems to be dependent upon the intent and the item-writing practices of the item constructor. Items susceptible to effective use of the convergence principle are the result of a particular approach to writing distractors. This approach involves determining the correct response first and then determining two characteristics of that response. Incorrect responses are generated by finding other instances of each of those characteristics (but not both in the same incorrect option). The test-wise student can use his/her deductive reasoning ability on this type of item if he/she is aware that the item writer uses

this approach to write test items. Smith (1982) found that college students (N=26) were able to use the convergence strategy. However, the test he used consisted of 20 four-choice multiple-choice items with no item stems and with all of the items susceptible to the use of the convergence principle. The students in his study obtained a mean score of 7.4 (37.0%) which was statistically significant when compared to the chance level. The entering medical students obtained a mean score nearly identical to that obtained by Smith when the items measuring the convergence principle were imbedded within items measuring five other TW principles. After being taught the TW principles of convergence and length of keyed option, workshop participants were able to show significant gains. However, the mean score on the convergence subtest after this instruction was still only 2.67 (66.8%) while the other five mean subtest scores were all at or above 3.47 (86.8%).

Further development and testing of the MOOMI is currently underway with medical, nursing, and premedical students. Additionally, research on the long-term effects of training in test taking skills is being conducted with medical and nursing students.

REFERENCES

Dorland's Illustrated Medical Dictionary. Philadelphia: W. B. Saunders Company, 25th Edition, 1974.

Millman, J., Bishop, C. H., and Ebel, R. An analysis of test-wiseness. Educational and Psychological Measurement, 1965, 25, 707-726.

Sarnacki, R. E. An examination of test-wiseness in the cognitive domain. Review of Educational Research, 1979, 49, 252-279.

Sarnacki, R. E. The effects of test-wiseness in medical education. Evaluation and the Health Professions, 1981, 4, 207-221.

Smith, J. K. Converging on correct answers: a peculiarity of multiple choice items. Journal of Educational Measurement, 1982, 19, 211-220.

TABLE 1

Means, standard deviations, standard errors of measurement, and Kuder-Richardson 20 reliability values for the subtest and total test scores (Forms A and B) for the Measure of Obscure Medical Information (N=39)

	Form A				Form B				Combined			
	Mean	SD	SEM	KR-20	Mean	SD	SEM	KR-20	Mean	SD	SEM	KR-20
Similar Options	2.82	0.87	0.85	0.05	2.74	0.95	0.82	0.25	5.56	1.55	1.16	0.44
Umbrella Term	2.79	1.11	0.82	0.46	3.28	0.85	0.63	0.44	6.08	1.67	1.04	0.61
Give-Aways	3.26	0.90	0.70	0.39	3.41	0.67	0.66	0.01	6.67	1.27	0.97	0.42
Convergence	1.33	1.05	0.81	0.40	1.64	1.07	0.94	0.24	2.97	1.80	1.22	0.54
Length	1.59	1.01	0.90	0.19	1.97	1.25	0.85	0.53	3.56	1.91	1.24	0.57
Stem-Option Cue	2.82	0.90	0.87	0.06	3.49	0.78	0.62	0.38	6.31	1.42	1.05	0.45
TOTAL TEST	14.62	3.56	2.04	0.67	16.54	3.05	1.94	0.60	31.15	6.14	2.81	0.79

TABLE 2

Alternate form reliability coefficients for subtest and total test scores (Forms A and B) for the Measure of Obscure Medical Information (N=39)

Similar Options	Umbrella Term	Give-Aways	Convergence	Length	Stem-Option Cue	TOTAL
0.44	0.44	0.29	0.45	0.42	0.42	0.72

TABLE 3

Pearson product-moment correlation coefficients among the subtest and total test scores (Forms A and B combined) on the Measure of Obscure Medical Information (N=39)

	Similar Options	Umbrella Term	Give-Aways	Convergence	Length	Stem-Option Cue
Umbrella Term	0.47					
Give-Aways	0.40	0.41				
Convergence	0.16	0.18	0.03			
Length	0.38	0.33	-0.01	0.09		
Stem-Option Cue	0.40	0.48	0.33	0.25	0.51	
TOTAL	0.72	0.74	0.50	0.48	0.64	0.76

TABLE 4

Pretest and posttest means, standard deviations, changes, and correlations for the subtest and total test scores for the Measure of Obscure Medical Information (N=15)

	Pretest (Form A)		Posttest (Form B)		Change		Pretest/ Posttest	t
	Mean	SD	Mean	SD	Mean	SD	Correlation	Value
Similar Options	2.53	1.36	3.47	0.99	0.94	1.58	.12	2.30*
Umbrella Term	3.13	1.13	4.00	0.00	0.87	1.13	-	2.98*
Give-Aways	3.40	0.99	3.60	0.63	0.20	0.86	.50	0.90
Convergence	0.87	0.99	2.67	0.98	1.80	1.57	-.27	4.44*
Length	2.20	1.21	3.80	0.56	1.60	1.18	.02	5.25*
Stem-Option Cue	3.00	0.65	3.53	0.64	0.53	0.64	.51	3.21*
TOTAL TEST	15.13	3.04	21.07	2.46	5.94	3.37	.27	6.83*

*Significant at $p < .05$

TABLE 5

Means, standard deviations, standard errors of measurement, and Kuder-Richardson 20 reliability values for the subtest and total test scores (Forms A and B) for the revision of the Measure of Obscure Medical Information (N=39)

	Form A				Form B				Combined			
	Mean	SD	SEM	KR-20	Mean	SD	SEM	KR-20	Mean	SD	SEM	KR-20
Similar Options	2.82	0.87	0.85	0.05	2.74	0.96	0.83	0.25	5.56	1.55	1.16	0.44
Umbrella Term	3.00	0.99	0.76	0.41	3.08	0.89	0.76	0.28	6.08	1.67	1.04	0.61
Give-Aways	3.33	0.80	0.68	0.27	3.33	0.69	0.69	0.00	6.67	1.27	0.97	0.42
Convergence	1.33	1.05	0.81	0.40	1.64	1.07	0.94	0.24	2.97	1.80	1.22	0.54
Length	1.72	0.99	0.95	0.08	1.85	1.23	0.83	0.55	3.56	1.91	1.24	0.57
Stem-Option Cue	3.18	0.78	0.77	0.02	3.13	0.99	0.70	0.50	6.31	1.42	1.05	0.45
TOTAL TEST	15.38	3.33	1.97	0.65	15.77	3.38	2.00	0.65	31.15	6.14	2.81	0.79

TABLE 6

Alternate form reliability coefficients for subtest and total test scores (Forms A and B) for the revision of the Measure of Obscure Medical Information (N=39)

Similar Options	Umbrella Term	Give-Aways	Convergence	Length	Stem-Option Cue	TOTAL
0.44	0.58	0.45	0.45	0.47	0.27	0.68